ROBUST AUTOMATIC SPEECH RECOGNITION IN REVERBERANT ENVIRONMENTS BY MODEL SELECTION

Laurent Couvreur

Signal Processing Department
Faculté Polytechnique de Mons
B-7000 Mons, Belgium
e-mail: lcouv@tcts.fpms.ac.be

Christophe Couvreur

Corporate R&D
Lernout & Hauspie Speech Products
B-1780 Wemmel, Belgium
e-mail: christophe.couvreur@lhs.be

ABSTRACT

This paper presents a method for robust automatic speech recognition (ASR) in reverberant environments. Our approach consists in the selection during operation of an acoustic model out of a library of models trained in various reverberant conditions. The best model is selected by blindly estimating the full-band reverberation time. The estimation procedure is entirely based on the short-term log-energy sequence of the utterance to be recognized. Speech recognition experiments in simulated and real reverberant environments show the efficiency of our approach which outperforms standard channel normalization techniques.

1. INTRODUCTION

Automatic speech recognition is a key component in hands-free man-machine interaction. State-of-the-art speech recognizers are based on statistical acoustic models which are commonly trained on *clean* material, i.e. noise-free and echo-free speech. In many applications, speech recognizers are deployed in reverberant enclosures, and the distance between the speaker and the microphone is generally higher than the so-called critical distance [9]. That is, most of the acoustic energy reaches the microphone after one or more reflections. The speech signal can be highly distorted by this room reverberation. Consequently, the performance of recognizers trained on clean speech deteriorates severely in reverberant environments because of the mismatch between the training and the operating conditions.

In [3], we showed that training acoustic models on artificially reverberated speech can provide robust models for the recognition of distant-talking speech in reverberant environments. Reverberated training material can be generated by convolving clean speech with room impulse responses. Instead of using measured room impulse responses [6, 11], we proposed in [3] to produce reverberated speech by processing clean speech with a filter whose finite-length impulse response is designed to match a high-level, perceptually meaningful, acoustic property of the target reverberant operating environment: the reverberation time T_{60}. In practice, the reverberation time of the operating environment is generally unknown. One can suggest to build a single acoustic model by multi-style training to account for various reverberation times. However, better performance is achieved if multiple acoustic models are trained separately for different reverberant conditions, and the best model is selected during operation [3]. The best model is the model trained in the reverberant conditions most closely matching that of the operating environment. In this paper, we propose an algorithm for blindly estimating the reverberation time of a room from speech signal recorded in that room. Once the reverberation time of the room has been estimated, the model with the closest reverberation time can be selected in a library of off-line trained reverberated acoustic models.

This paper is organized as follows. In the next section, we briefly review the procedure for training acoustic models on artificially reverberated speech. Section 3 presents the model selection procedure. Experimental results for recognition of connected digit sequences in reverberant environments are reported in section 4. Conclusions are drawn in section 5.

2. TRAINING PROCEDURE

We assume that the effect of room reverberation on a speech recognizer can be entirely characterized by the reverberation time T_{60}. It is expressed in seconds and defined as the time interval in which the sound energy in a room reaches one millionth of its initial value (-60dB) after interrupting the sound source. We further assume that the sound field is diffuse and that the reverberation time is frequency independent. Under those hypotheses, room reverberation can be rendered by convolving clean speech with a synthetic impulse response h_n. The impulse response h_n can be obtained by shaping a Gaussian white random sequence with a decaying exponential whose damping constant is directly related to the reverberation time. The detailed description of the synthetic impulse response computation can be found in [3].

Once a reverberated database has been generated by convolving a clean speech database with h_n, an acoustic model corresponding to the specified T_{60} can be trained. Note that h_n is recomputed several times during the generation of the reverberated database for "smoothing" the trained model.

Repeating the process for different values of T_{60}, a library of acoustic models can be build for various reverberation times.

3. SELECTION PROCEDURE

During operation, we want to select the best acoustic model. The best model is the model trained in the reverberant

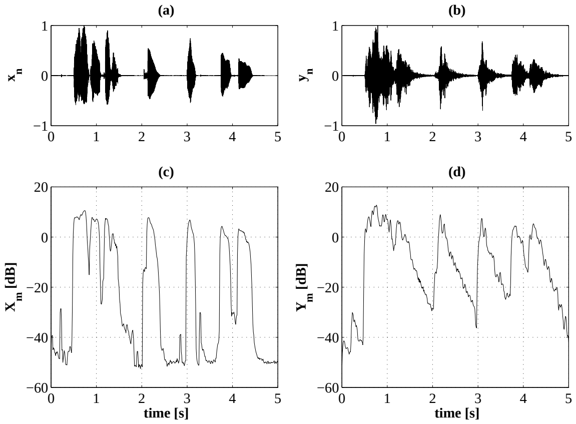

Figure 1: Waveforms for (a) a clean speech utterance x_n and (b) its reverberant version y_n, and the corresponding L_{eq} sequences (c) X_m and (d) Y_m for T_w=30ms and F_r=100Hz.

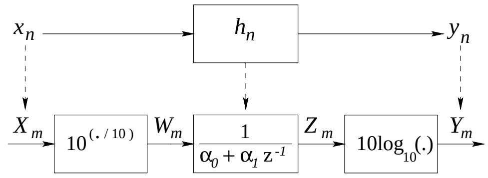

Figure 2: Room reverberation process (upper part) for temporal signal and equivalent diffuse model (lower part) for L_{eq} sequence.

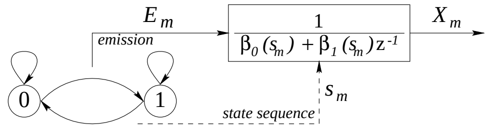

Figure 3: 2-states one-dimensional first-order LP-HMM for modeling L_{eq} sequence of clean speech (0: silence state, 1: speech state).

conditions (characterized by the reverberation time) most closely matching that of the operating reverberant environment. Hence, the problem of selecting a model reduces to estimating T_{60} from the utterances to be recognized. In this section, we outline an algorithm for blindly estimating T_{60} from samples of reverberated speech. The algebraic details can be found in [4].

3.1. Room Reverberation Model

The most detailed model of room reverberation is the room impulse response between the speaker and the microphone. One can propose to identify blindly the room impulse response from recorded reverberated speech, and then compute the reverberation time from the estimated impulse response, e.g. using Schroeder's method [9]. Since the blind identification of a room impulse response is a sensitive task, we propose to use a simpler model of room reverberation in order to simplify the T_{60} estimation problem. We decide to model the impact of room reverberation on the short-term log-energy (L_{eq}) sequence X_m instead of on the clean speech signal x_n,

$$X_m \triangleq 10\log_{10}\left(\frac{1}{N_w}\sum_{n=mN_r}^{mN_r+N_w-1} x_n^2\right), \tag{1}$$

with $N_w \triangleq T_w \times F_s$ and $N_r \triangleq F_s/F_r$, where T_w, F_s and F_r denote the analysis frame length [s], the sampling frequency [Hz] and the frame rate [Hz], respectively. Figure 1 gives an example of a clean speech utterance x_n and its reverberated version y_n obtained by convolving x_n with a typical room impulse response h_n. The figure also shows the distortion on the corresponding L_{eq} sequences X_m and Y_m computed after proper normalization of the speech signals. For pure diffuse sound fields, the decays of Y_m from peak to valley should be exactly linear, and thus exponential in the linear energy domain. That is, the impact of room reverberation can be modeled by a first order AR filter,

$$W_m = \alpha_0 Z_m + \alpha_1 Z_{m-1} \tag{2}$$

where $W_m \triangleq 10^{X_m/10}$ and $Z_m \triangleq 10^{Y_m/10}$ denote the short-term linear energy sequences of the clean and reverberated speech signals, respectively. The assumed reverberation model is summarized in figure 2. In the sequel, we describe a method for Maximum Likelihood (ML) estimation of the AR coefficients (α_0, α_1) from an observation of the L_{eq} sequence Y_m only. Our estimation algorithm requires a statistical model for the echo-free L_{eq} sequence X_m which is briefly presented in the next section. Once α_1 has been estimated, T_{60} can be derived via [4],

$$T_{60} = \log 10^6/(-\log(-\alpha_1) \times F_r). \tag{3}$$

3.2. L_{eq} Source Model

The clean speech L_{eq} sequence X_m is typically nonstationary and characterized by two states, called the silence and speech states. Furthermore, successive values are undoubtedly not statistically independent: they are correlated (see figure 1.c). Hence, we choose to model X_m by a 2-states one-dimensional Linear Predictive Hidden Markov Model (LP-HMM) [8]. In this model, the L_{eq} sequence X_m is obtained by processing the emission sequence E_m with an AR filter of order K,

$$\beta_0(s_m)X_m = E_m - \sum_{k=1}^{K}\beta_k(s_m)X_{m-k} \tag{4}$$

whose coefficients $\beta_k(s_m)$, $k = 0, \ldots, K$, are function of the HMM state sequence s_m. The emissions E_m are assumed to be conditionally independent given the state sequence s_m and have a Gaussian distribution with mean μ_i and variance σ_i for $s_m = i$, $i = 0, 1$. To complete our model, we define the transition probabilities $a_{ij} \triangleq P[s_m = j | s_{m-1} = i]$. All the parameters can be estimated by an Expectation-Maximization (EM) algorithm [8] from L_{eq} sequences extracted from a clean speech database. Figure 3 illustrates a 2-states one-dimensional LP-HMM with AR filters limited to first order ($K = 1$) which is used in this work. Table 1 gives the parameters of the model trained on a clean part of the AURORA speech database [2].

Table 1: Parameters of a 2-states one-dimensional first-order LP-HMM for L_{eq} sequence of clean speech for T_w=30ms and F_r=100Hz.

$s_m = i$	a_{ii}	a_{ij}	μ_i	σ_i	$\beta_0(i)$	$\beta_1(i)$
0	0.95	0.05	-4.3	4.2	1.0	-0.92
1	0.03	0.97	1.1	3.2	1.0	-0.77

3.3. T_{60} Estimation Algorithm

In this section, we describe a method for blindly estimating T_{60}. The algorithm is based on ML stochastic matching [10]. Given a statistical model of the unobserved clean L_{eq} sequence X_m (section 3.2), the parameters (α_0, α_1) of the distortion model (section 3.1) are estimated so as to maximize the likelihood of the observed reverberated L_{eq} sequence Y_m. Maximization of the likelihood with respect to (α_0, α_1) is performed via the EM algorithm. Given an observed L_{eq} sequence Y_0^M of length $M+1$ and current estimates $(\alpha_0^{(\ell)}, \alpha_1^{(\ell)})$, we first compute (E-step) the auxiliary function,

$$\begin{aligned} & Q(\alpha_0^{(\ell+1)}, \alpha_1^{(\ell+1)} | \alpha_0^{(\ell)}, \alpha_1^{(\ell)}) \\ \triangleq & E\left[\log p(Y_0^M, s_0^M | \alpha_0^{(\ell+1)}, \alpha_1^{(\ell+1)}) \;\middle|\; Y_0^M, \alpha_0^{(\ell)}, \alpha_1^{(\ell)}\right] \end{aligned} \tag{5}$$

where s_0^M denotes the hidden state sequence of the LP-HMM. We then find (M-step) closed-form re-estimation formulae by setting the first derivatives of (5) with respect to $(\alpha_0^{(\ell+1)}, \alpha_1^{(\ell+1)})$ to zero. The resulting iterative estimation algorithm is outlined below:

1. Initialize the estimates of the distortion parameters, $(\alpha_0^{(0)}, \alpha_1^{(0)})$ and set $\ell = 0$;
2. Compute $Z_m = 10^{Y_m/10}$, $m = 0, \dots, M$, and apply the inverse filter $\alpha_0^{(\ell)} + \alpha_1^{(\ell)} z^{-1}$ to obtain $W_m^{(\ell)} = \alpha_0^{(\ell)} Z_m + \alpha_1^{(\ell)} Z_{m-1}$;
3. Estimate the *a posteriori* state probabilities $\gamma_m^{(\ell)}(i) \triangleq P[s_m^{(\ell)} = i | Y_0^M]$, $i = 0, 1$ via the *Forward-Backward* algorithm [8] given the LP-HMM parameters and $X_m^{(\ell)} = 10 \log_{10} W_m^{(\ell)}$, $m = 0, \dots, M$;
4. Apply the re-estimation formulae [4] based on the LP-HMM parameters, the *a posteriori* state probabilities $\gamma_m^{(\ell)}(i)$ and the observations Y_0^M, and obtain updated estimates of the distortion parameters $(\alpha_0^{(\ell+1)}, \alpha_1^{(\ell+1)})$;
5. Set $\ell = \ell + 1$ and go to 2 unless convergence is reached;
6. Derive T_{60} from $\alpha_1^{(\ell)}$ via (3).

Note that a Viterbi approximation may be used [10] for a fast implementation of the algorithm. In that case, only the most likely state sequence is retained to express the likelihood, i.e. the *a posteriori* probabilities $\gamma_m(i)$ are constrained to be equal to 0 or 1.

4. EXPERIMENTAL RESULTS

The speech corpus used in this work comes from the clean part of the AURORA [2] speech database and consists of connected digit sequences. The corpus is divided into a training set of 8840 utterances and a test set of 1001 utterances, pronounced by 110 speakers and 104 other speakers, respectively. Recognition experiments are performed with a phoneme-based hybrid Multilayer Perceptron(MLP)/HMM recognizer. The phoneme *a posteriori* probabilities are estimated by a MLP fed with acoustic features computed from 30ms long/10ms overlapping frames of signal sampled at 8kHz. Speech decoding is done by Viterbi search, without any pruning or grammar constraints.

Table 2: Performances of baseline recognizers with various front-ends for echo-free speech.

Front-end	WER [%]	SUB/DEL/INS [%]
MFCC	1.7	0.7/0.5/0.5
MFCC-CMS	1.8	0.7/0.6/0.5
logRASTA-PLP	1.9	0.7/0.6/0.6

4.1. Baseline Models

First, we trained acoustic models on the clean training set for three front-ends: Mel-warped frequency cepstral coefficients (MFCC), MFCC with cepstral mean subtraction (CMS) [5] and logRASTA-PLP [7] coefficients. The last two front-ends are known to be robust to channel distortion. We then used the resulting systems to recognize the clean test set. Table 2 gives the results of these baseline systems in terms of the word error rate[1] (WER). As expected, they all achieve satisfactorily performances for echo-free speech.

4.2. Reverberated Models

Next, we trained eight acoustic models on artificially reverberated training sets for T_{60} varying uniformly from 200ms to 1600ms. For each T_{60}, the corresponding training set was obtained by using the method depicted in section 2. Meanwhile, test sets were generated by convolving the clean test set with room impulse responses computed by the Image Method [1]. The wall absorption coefficients of the reverberant enclosure simulator were chosen to get specific reverberation times. Table 3 reports cross-testing results. We see that the lowest WER is always achieved by the acoustic model most closely matching the testing conditions (main diagonal). Even if there is no acoustic model which matches exactly the test T_{60}, the performance of the selected model does not degrade much if the grid for T_{60} in the library of acoustic models is tight enough. As could have been expected, WER increases for the matching acoustic model as the reverberation becomes stronger.

4.3. Model Selection Approach

Finally, we tested our model selection approach by blind estimation of T_{60}. Test sets were generated by mixing groups of utterances reverberated at different levels. Each group was at least 3s long and obtained by convolving clean utterances with a room impulse response corresponding to a specific T_{60}. Prior to its recognition, every group was processed: the L_{eq} sequence was computed for T_w = 30ms and F_r = 100Hz,

[1]Sum of the substitution (SUB), deletion (DEL) and insertion (INS) error rates.

Table 3: Performances WER [%] of MFCC-based acoustic models trained on artificially reverberated speech for various reverberant testing conditions.

Test set	Training set								
	clean	$T_{60} = 200$ms	400ms	600ms	800ms	1000ms	1200ms	1400ms	1600ms
clean	**1.7**	2.9	7.6	11.9	15.9	19.8	20.6	22.7	23.8
$T_{60} = 200$ms	7.0	**3.6**	4.5	6.4	9.8	12.5	13.7	15.1	16.1
300ms	7.8	**3.9**	**4.4**	6.4	9.8	12.3	13.9	15.0	15.7
400ms	18.7	9.6	**5.2**	5.7	8.5	12.2	12.8	14.7	15.3
500ms	20.1	11.2	**5.9**	**5.9**	8.7	12.2	12.8	14.7	15.4
600ms	29.7	20.2	11.3	**9.2**	10.0	12.6	13.7	15.6	16.4
700ms	33.2	24.7	14.9	**11.2**	**11.3**	13.6	14.4	16.1	17.1
800ms	41.0	33.7	22.1	17.3	**14.0**	15.9	16.6	18.2	19.1
1000ms	43.4	35.8	24.0	20.4	16.0	**17.0**	17.1	18.7	19.7
1200ms	49.3	43.1	32.0	27.9	20.9	20.7	**20.4**	21.6	22.1
1400ms	51.1	48.5	36.8	33.5	26.0	24.9	23.2	**24.5**	24.4
1600ms	52.9	50.1	37.3	36.6	28.1	26.7	25.3	25.1	**25.1**

Table 4: Comparison between performances WER (SUB/DEL/INS) [%] of two standard normalization techniques, our model selection method and the "Oracle" method.

Method	Setup	
	Test A	Test B
MFCC-CMS	35.9(10.8/15.0/10.1)	21.2(7.5/10.7/3.0)
logRASTA-PLP	35.4(12.7/14.0/8.7)	22.4(8.7/9.8/3.9)
Model Selection	15.6(5.5/6.4/3.7)	12.8(4.9/5.2/2.7)
"Oracle"	13.7(4.7/6.2/2.8)	—

T_{60} was estimated and the most closely matching MLP of the library was activated. Two sets of experiments were performed: test sets were generated by convolution with room impulse responses either computed in one of the previous simulated rooms (test A), or measured in real reverberant enclosures (test B). Table 4 shows that the proposed model selection method outperforms systems based on standard channel robust acoustic features. Furthermore, it approaches the performance of the "Oracle" method for which T_{60} is assumed to be known exactly (only for test A) and the best model is always selected.

5. CONCLUDING REMARKS

We have proposed an algorithm for blind estimation of the reverberation time, and successfully applied it for robust speech recognition in reverberant environments by acoustic model selection. Further improvements can be expected by relaxing the main hypothesis which supposes that the reverberation time is frequency independent. To do so, the method has to be extended to a multiband approach for which T_{60} is assumed constant inside frequency subbands only.

ACKNOWLEDGMENT

The authors would like to thank Prof. S. Nakamura from ATR and Dr. J. Hopgood from CUED for providing the real room impulse responses used in section 4.3 of this work.

REFERENCES

[1] J. B. Allen and D. A. Berkley, "Image Method for Efficiently Simulating Small-Room Acoustics", *Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[2] AURORA database - http://www.elda.fr/aurora2.html.

[3] L. Couvreur, C. Couvreur and C. Ris, "A Corpus-Based Approach for Robust ASR in Reverberant Environments", *Proc. of ICSLP'2000*, vol. 1, pp. 397–400, Beiging, China, Oct. 2000.

[4] L. Couvreur and C. Couvreur, "Blind Estimation of Acoustical Reverberation Times", *in preparation*.

[5] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[6] D. Giuliani, M. Matassoni, M. Omologo and P. Svaizer, "Training of HMM with Filtered Speech Material for Hands-free Recognition", *Proc. of ICASSP'99*, vol. 1, pp. 449–452, Phoenix, USA, Mar. 1999.

[7] H. Hermansky and N. Morgan, "RASTA Processing of Speech", *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[8] P. Kenny, M. Lennig and P. Mermelstein, "A Linear Predictive HMM for Vector-Valued Observations with Applications to Speech Recognition", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 38, no. 2, pp. 220–225, Feb. 1990.

[9] H. Kuttruff, *Room Acoustics*, Elsevier Applied Science, 3rd edition, 1991.

[10] A. Sankar and C.-H. Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition", *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, May 1996.

[11] Y. Shimizu, S. Kajita, K. Takeda and F. Itakura, "Speech Recognition Based on Space Diversity Using Distributed Multi-Microphone", *Proc. of ICASSP'2000*, vol. 3, pp. 1747–1750, Istanbul, Turkey, Jun. 2000.